Supplement to "A Tuning-free Robust and Efficient Approach to High-dimensional Regression"

Lan Wang, Bo Peng, Jelena Bradic, Runze Li and Yunan Wu

1 Additional technical results for proving Theorems 1& 2

Lemma A. *Assume $X_1, \dots, X_n$ are mutually independent random variables. Let $U_r(X_1, \dots, X_n) = \binom{n}{r}^{-1} \sum_{i_1 < \dots < i_r} h(X_{i_1}, \dots, X_{i_r})$ where $h(\cdot)$ is a symmetric kernel function $h(\cdot)$ and $1 \le r \le n$ is a fixed positive integer. Assume that there exists a constant $M > 0$ such that $|h(X_{i_1}, \dots, X_{i_r})| \le M$. Then for any $t \ge 0$,*

$$P\Big(\big|U_r(X_1, \dots, X_n) - \theta(h)\big| > t\Big) \le 2\exp\Big(-\frac{nt^2}{8M^2}\Big),$$

where $\theta(h) = E\big(h(X_{i_1}, \dots, X_{i_r})\big)$.

Proof of Lemma A. This result follows from the bounded difference inequality.

[1]Lan Wang is Professor, Department of Management Science, University of Miami. Emails: lanwang@mbs.miami.edu. Bo Peng is a data scientist at Adobe. Email: bpeng@adobe.com. Jelena Bradic is Associate Professor, Department of Mathematics and Halicioglu Data Science Institute, University of California at San Diego. Email: jbradic@ucsd.edu. Runze Li is Eberly Family Chair Professor, Department of Statistics, Pennsylvania State University. Email: rzli@psu.edu. Yunan Wu is graduate student, School of Statistics, University of Minnesota. Emails: wuxx1375@umn.edu. Wang and Wu's research was supported by NSF DMS-1712706, NSF OAC-1940160 and FRGMS-1952373. Bradic's research was supported by NSF DMS-1712481. Li's research was supported by NIDA grant P50 DA039838, NSF DMS 1512422 and DMS 1820702. The authors are indebted to the referees, the associate editor and the Co-editors for their valuable comments, which have significantly improved the paper.

Proof of Lemma 2. (i) Denote $\widehat{\boldsymbol{\gamma}}(\lambda)$ by $\widehat{\boldsymbol{\gamma}}$ for simplicity. By the definition of $\widehat{\boldsymbol{\beta}}(\lambda)$, we have, $Q_n(\widehat{\boldsymbol{\gamma}}) + \lambda||\widehat{\boldsymbol{\beta}}||_1 \leq Q_n(\mathbf{0}) + \lambda||\boldsymbol{\beta}_0||_1$. This implies

$$Q_n(\widehat{\boldsymbol{\gamma}}) - Q_n(\mathbf{0}) \leq \lambda\big(||\boldsymbol{\beta}_0||_1 - ||\widehat{\boldsymbol{\beta}}||_1\big) \leq \lambda\big(||\widehat{\boldsymbol{\gamma}}_A||_1 - ||\widehat{\boldsymbol{\gamma}}_{A^c}||_1\big). \qquad (1)$$

By the convexity of Q_n and the definition of subdifferential,

$$\begin{aligned} Q_n(\widehat{\boldsymbol{\gamma}}) - Q_n(\mathbf{0}) \geq -2\big[n(n-1)\big]^{-1}\mathbf{X}^T\boldsymbol{\xi}\widehat{\boldsymbol{\gamma}} &\geq -2\big[n(n-1)\big]^{-1}||\widehat{\boldsymbol{\gamma}}||_1||\mathbf{X}^T\boldsymbol{\xi}||_\infty \\ &\geq -\frac{\lambda}{c}(||\widehat{\boldsymbol{\gamma}}_A||_1 + ||\widehat{\boldsymbol{\gamma}}_{A^c}||_1), \qquad (2) \end{aligned}$$

(1) and (2) together imply $||\widehat{\boldsymbol{\gamma}}_{A^c}||_1 \leq \bar{c}||\widehat{\boldsymbol{\gamma}}_A||_1$.
(ii) Write $\mathbf{S}_n = (s_1, \ldots, s_p)^T$, where $s_k = \big[n(n-1)\big]^{-1}\sum\sum_{i\neq j}(x_{jk} - x_{ik})\text{sign}(\epsilon_i - \epsilon_j)$, $k = 1\ldots, p$. By the union bound, for $c_0 = 4\sqrt{2}b_1 c$,

$$P\big(c||\mathbf{S}_n||_\infty \geq lc_0\sqrt{\log p/n}\big) \leq \sum_{k=1}^{p} P\big(|s_k| \geq c^{-1}lc_0\sqrt{\log p/n}\big).$$

For each s_k, we apply the concentration inequality for U-statistics (Lemma A) and obtain

$$P\big(|s_k| \geq c^{-1}lc_0\sqrt{\log p/n}\big) \leq 2\exp\Big(-\frac{l^2c_0^2\log p}{32b_1^2c^2}\Big) \leq 2\exp\big(-l^2\log p\big).$$

Thus $P\big(c||\mathbf{S}_n||_\infty < lc_0\sqrt{\log p/n}\big) \geq 1 - 2\exp(-(l^2-1)\log p)$. (iii) Taking $l = 2$ in (ii), we have $P\big(c||\mathbf{S}_n||_\infty > 2c_0\sqrt{\log p/n}\big) \leq \alpha_0$. The conclusion follows by the definition of quantile. □

Proof of Lemma 3. Let

$$G_n(\mathbf{u}) = n^{-1}q^{-1}[L_n(\boldsymbol{\beta}_{01} + n^{-1/2}q^{1/2}\mathbf{u}) - L_n(\boldsymbol{\beta}_{01})],$$

where $L_n(\boldsymbol{\beta}_1) = \sum\sum_{i\neq j}|(\epsilon_i - \epsilon_j) - (\mathbf{x}_{1i} - \mathbf{x}_{1j})^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01})|$. It is sufficient to show that $\forall\ \eta > 0$, there exists a $\Delta > 0$ such that

$$P\Big(\inf_{\mathbf{u}\in\mathcal{R}^q, ||\mathbf{u}||_2=\Delta} L_n(\boldsymbol{\beta}_{01} + n^{-1/2}q^{1/2}\mathbf{u}) > L_n(\boldsymbol{\beta}_{01})\Big) \geq 1 - \eta. \qquad (3)$$

By similar arguments as in the proof of Theorem 1, if (3) holds, then the convexity of $L_n(\boldsymbol{\beta}_1)$ implies that $\widehat{\boldsymbol{\beta}}_1$ is within the L_2-ball

$$\{\boldsymbol{\beta}_1 \in \mathcal{R}^q : ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}||_2 \leq \Delta n^{-1/2}q^{1/2}\}$$

with probability at least $1-\eta$. By Knight's identity (Koenker, 2005),

$$\begin{aligned} G_n(\mathbf{u}) &= 2n^{-1}q^{-1}\sum\sum_{i\neq j} n^{-1/2}q^{1/2}(\mathbf{x}_{1i}-\mathbf{x}_{1j})^T\mathbf{u}\big[I(\epsilon_i-\epsilon_j<0)-1/2\big] \\ &\quad +2n^{-1}q^{-1}\sum\sum_{i\neq j}\int_0^{n^{-1/2}q^{1/2}(\mathbf{x}_{1i}-\mathbf{x}_{1j})^T\mathbf{u}}\big[I(\epsilon_i-\epsilon_j<s)-I(\epsilon_i-\epsilon_j<0)\big]ds. \end{aligned}$$

Then for some ξ_{ij} between 0 and $n^{-1/2}q^{1/2}(\mathbf{x}_{1i}-\mathbf{x}_{1j})^T\mathbf{u}$, $i\neq j$, we have

$$\begin{aligned} \mathrm{E}(G_n(\mathbf{u})) &= 2n^{-1}q^{-1}\sum\sum_{i\neq j}\int_0^{n^{-1/2}q^{1/2}(\mathbf{x}_{1i}-\mathbf{x}_{1j})^T\mathbf{u}}\big[F^*(s)-F^*(0)\big]ds \\ &= 2n^{-1}q^{-1}\sum\sum_{i\neq j}\int_0^{n^{-1/2}q^{1/2}(\mathbf{x}_{1i}-\mathbf{x}_{1j})^T\mathbf{u}} f^*(\xi_{ij})sds \\ &\geq cn^{-2}\sum\sum_{i\neq j}\big[(\mathbf{x}_{1i}-\mathbf{x}_{1j})^T\mathbf{u}\big]^2 \geq c\lambda_{\min}\big(n^{-1}\sum_{i=1}^n \mathbf{x}_{1i}\mathbf{x}_{1i}^T\big)||\mathbf{u}||_2^2 \\ &\geq c\Delta^2, \end{aligned}$$

where the second last inequality uses condition (C1). Similarly as the proof of Theorem 1, by McDiarmid's inequality, for any $\mathbf{u}$ such that $||\mathbf{u}||_2=\Delta$ and $\forall\ t>0$,

$$P\Big(|G_n(\mathbf{u})-E(G_n(\mathbf{u}))|\geq t\Big)\leq 2\exp\Big\{-c\frac{t^2}{\Delta^2}\Big\}$$

for some constant $c>0$. Hence, for Δ sufficiently large,

$$P\Big(|G_n(\mathbf{u})-E(G_n(\mathbf{u})\Big)|\leq\Delta^{3/2})\geq 1-\eta. \tag{4}$$

This holds uniformly in $\mathbf{u}$ in a bounded region since $G_n(\cdot)$ is convex (Pollard, 1991). The lemma holds because $\mathrm{E}(G_n(\mathbf{u}))$, which is positive and quadratic in Δ, dominates $\sup_{\mathbf{u}\in\mathcal{R}^q,||\mathbf{u}||_2=\Delta}|G_n(\mathbf{u})-E(G_n(\mathbf{u}))|$ for sufficiently large Δ. $\square$

Lemma B. *Assume conditions of Theorem 2 of the main paper are satisfied. There exist v_{ij}^* which satisfies $v_{ij}^*=0$ if $Y_i-Y_j\neq(\mathbf{x}_i-\mathbf{x}_j)^T\widehat{\boldsymbol{\beta}}^{(o)}$ and $v_{ij}^*\in$*

$[-1,1]$ *if* $Y_i - Y_j = (\mathbf{x}_i - \mathbf{x}_j)^T \widehat{\boldsymbol{\beta}}^{(o)}$*, such that for* $\delta_k(\widehat{\boldsymbol{\beta}}^{(o)})$ *with* $v_{ij} = v_{ij}^*$*, with probability approaching one, we have*

$$
\begin{aligned}
\delta_k(\widehat{\boldsymbol{\beta}}^{(o)}) &= 0, \ k = 1, \ldots, q. && (5)\\
|\delta_k(\widehat{\boldsymbol{\beta}}^{(o)})| &< a_1\eta, \ k = q+1, \ldots, p. && (6)
\end{aligned}
$$

Proof of Lemma B. Equality (5) follows from the subgradient condition for convex optimization. To prove (6), we first note that with probability one the number of (i,j) pairs such that $Y_i - Y_j = (\mathbf{x}_i - \mathbf{x}_j)^T \widehat{\boldsymbol{\beta}}^{(o)}$ is of order $O(q)$, see for example Section 2.2 of Koenker (2005). Hence

$$
\max_{q+1 \le k \le p} \Big| \big[n(n-1)\big]^{-1} \sum\sum_{i \ne j} (x_{jk} - x_{ik}) v_{ij}^* I\big(Y_i - Y_j = (\mathbf{x}_i - \mathbf{x}_j)^T \widehat{\boldsymbol{\beta}}^{(o)}\big) \Big| = o(\eta).
$$

To prove (6), it suffices to show

$$
\begin{aligned}
&P\Big(\big[n(n-1)\big]^{-1} \Big| \sum\sum_{i \ne j} (x_{jk} - x_{ik}) \mathrm{sign}\big(Y_i - Y_j - (\mathbf{x}_i - \mathbf{x}_j)^T \widehat{\boldsymbol{\beta}}^{(o)}\big) \Big| > a_1\eta, \\
&\qquad \text{for some } k = q+1, \ldots, p\Big) \to 0.
\end{aligned}
$$

The left-hand side of the above expression is bounded above by

$$
\begin{aligned}
& P\Big(\max_{q+1 \le k \le p} \Big| \sum\sum_{i \ne j} (x_{jk} - x_{ik}) \big[I\big(Y_i - Y_j - (\mathbf{x}_i - \mathbf{x}_j)^T \widehat{\boldsymbol{\beta}}^{(o)} > 0\big) - 1/2 \big] \Big| \\
& \qquad > 2a_1 n(n-1)\eta \Big) \\
\le\ & P\Big(\max_{q+1 \le k \le p} \Big| \sum\sum_{i \ne j} (x_{jk} - x_{ik}) \big[I\big(Y_i - Y_j - (\mathbf{x}_i - \mathbf{x}_j)^T \widehat{\boldsymbol{\beta}}^{(o)} > 0\big) \\
& \qquad - I\big(Y_i - Y_j - (\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\beta}_0 > 0\big) \big] \Big| > a_1 n(n-1)\eta \Big) \\
& + P\Big(\max_{q+1 \le k \le p} \Big| \sum\sum_{i \ne j} (x_{jk} - x_{ik}) \big[I\big(\epsilon_i - \epsilon_j > 0\big) - 1/2 \big] \Big| > a_1 n(n-1)\eta \Big).
\end{aligned}
$$

Both terms go to zero, where the second probability goes to zero by applying Hoeffding's inequality for U-statistics and the first probability goes to zero by Lemma C below. $\square$

Lemma C. *Assume conditions of Theorem 2 are satisfied.*

$$P\Big(\max_{q+1\le k\le p}\Big|\sum\sum_{i\neq j}(x_{jk}-x_{ik})\big[I(Y_i-Y_j-(\mathbf{x}_i-\mathbf{x}_j)^T\widehat{\boldsymbol{\beta}}^{(o)}>0) \\ -I(Y_i-Y_j-(\mathbf{x}_i-\mathbf{x}_j)^T\boldsymbol{\beta}_0>0)\big]\Big|>a_1n(n-1)\eta\Big)\to 0. \quad (7)$$

Proof of Lemma C. For an arbitrary $\boldsymbol{\beta}\in\mathcal{R}^p$, let $\boldsymbol{\gamma}\in\mathcal{R}^q$ be the subvector that consists of the first q entries of $\boldsymbol{\beta}-\boldsymbol{\beta}_0$. Let $\zeta_{ij}=\epsilon_i-\epsilon_j$, $\widetilde{\mathbf{x}}_{1ij}=\mathbf{x}_{1i}-\mathbf{x}_{1j}$ where $\mathbf{x}_{1i}$ be the subvector that consists of the first q entries of $\mathbf{x}_i$ and $\mathbf{x}_{1i}$ is defined similarly; and let $\widetilde{x}_{ijk}=x_{ik}-x_{jk}$.

Because $\widehat{\boldsymbol{\beta}}^{(o)}$ is the oracle estimator, for any given positive constant Δ, we consider

$$\begin{aligned}
& P\Big(\max_{q+1\le k\le p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10}||_2\le\Delta\sqrt{q/n}}\Big|\sum\sum_{i\neq j}(x_{jk}-x_{ik})\big[I(Y_i-Y_j-(\mathbf{x}_{1i}-\mathbf{x}_{1j})^T\boldsymbol{\beta}_1>0) \\
&\qquad -I(Y_i-Y_j-(\mathbf{x}_{1i}-\mathbf{x}_{1j})^T\boldsymbol{\beta}_{10}>0)\big]\Big|>a_1n(n-1)\eta\Big) \\
\le\ & P\Big(\max_{q+1\le k\le p}\sup_{||\boldsymbol{\gamma}||_2\le\Delta\sqrt{q/n}}\Big|\sum\sum_{i\neq j}\widetilde{x}_{jik}\big[I(\zeta_{ij}>\widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma})-I(\zeta_{ij}>0) \\
&\qquad -P(\zeta_{ij}>\widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma})+P(\zeta_{ij}>0)\big]\Big|>a_1n(n-1)\eta/2\Big) \\
& +P\Big(\max_{q+1\le k\le p}\sup_{||\boldsymbol{\gamma}||_2\le\Delta\sqrt{q/n}}\Big|\sum\sum_{i\neq j}\widetilde{x}_{jik}\big[P(\zeta_{ij}>\widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma})-P(\zeta_{ij}>0)\big]\Big| \\
&\qquad >a_1n(n-1)\eta/2\Big) \\
=\ & D_1+D_2,
\end{aligned}$$

where the definition of D_i, $i=1,2$, is clear from the context. To see $D_2\to 0$ as $n\to\infty$, we observe that

$$\begin{aligned}
& \max_{q+1\le k\le p}\sup_{||\boldsymbol{\gamma}||_2\le\Delta\sqrt{q/n}}\Big|\sum\sum_{i\neq j}\widetilde{x}_{jik}\big[P(\zeta_{ij}>\widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma})-P(\zeta_{ij}>0)\big]\Big| \\
\le\ & c\max_{q+1\le k\le p}\sup_{||\boldsymbol{\gamma}||_2\le\Delta\sqrt{q/n}}\sum\sum_{i\neq j}|\widetilde{\mathbf{x}}_{ij}^T\boldsymbol{\gamma}| \\
\le\ & O(n^2)O(\sqrt{q/n})O(\sqrt{q})=O(qn^{3/2})=o(n^2\eta),
\end{aligned}$$

where c is some positive constant, by conditions (C1)–(C3) and the assumption of the Lemma.

To prove $D_1 \to 0$, we cover the ball $\big\{\boldsymbol{\gamma} \in \mathcal{R}^q : ||\boldsymbol{\gamma}||_2 \le \Delta\sqrt{q/n}\big\}$ with a net of balls with radius $\Delta\sqrt{q/n^5}$. This net can be constructed with cardinality $N \le d \cdot n^{4q_n}$ for some constant $d > 0$. Denote the N balls by $B(\mathbf{t}_1), \ldots, B(\mathbf{t}_N)$, where the ball $B(\mathbf{t}_r)$ is centered at $\mathbf{t}_r$, $r = 1, \ldots, N$. Then

$$
\begin{aligned}
& P\Big(\max_{q+1\le k\le p} \sup_{||\boldsymbol{\gamma}||_2 \le \Delta\sqrt{q/n}} \Big| \sum\sum_{i\ne j} \widetilde{x}_{jik}\big[I(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma}) - I(\zeta_{ij} > 0) \\
& \qquad - P(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma}) + P(\zeta_{ij} > 0)\big]\Big| > a_1 n(n-1)\eta/2\Big) \\
\le\; & \sum_{r=1}^N P\Big(\max_{q+1\le k\le p} \Big| \sum\sum_{i\ne j} \widetilde{x}_{jik}\big[I(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r) - I(\zeta_{ij} > 0) - P(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r) \\
& \qquad + P(\zeta_{ij} > 0)\big]\Big| > a_1 n(n-1)\eta/4\Big) \\
& + \sum_{r=1}^N P\Big(\max_{q+1\le k\le p} \sup_{||\boldsymbol{\gamma}-\mathbf{t}_r||_2 \le \Delta\sqrt{q/n^5}} \Big| \sum\sum_{i\ne j} \widetilde{x}_{jik}\big[I(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma}) - I(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r) \\
& \qquad - P(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma}) + P(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r)\big]\Big| > a_1 n(n-1)\eta/4\Big) \\
=\; & J_1 + J_2,
\end{aligned}
$$

where the definition of J_i, $i = 1, 2$, is clear from the context. To evaluate J_1, we have

$$
\begin{aligned}
J_1 \;\le\; & \sum_{r=1}^N \sum_{k=q+1}^p P\Big(\Big| \sum\sum_{i\ne j} \widetilde{x}_{jik}\big[I(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r) - I(\zeta_{ij} > 0) - P(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r) \\
& \qquad + P(\zeta_{ij} > 0)\big]\Big| > a_1 n(n-1)\eta/4\Big) \\
\le\; & 2\sum_{r=1}^N \sum_{k=q+1}^p \exp\big(-cn\eta^2\big) \\
\le\; & 2\exp\big(\log(N) + \log(p) - cn\eta^2\big) = o(1),
\end{aligned}
$$

where c is some positive constant, and the second inequality applies Hoeffding's inequality for U-statistics.

To evaluate J_2, we note that

$$
\begin{aligned}
& \widetilde{x}_{jik}\Big[I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma}\big) - I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r\big) - P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma}\big) + P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r\big)\Big] \\
\leq\ & 2b_1\Big[I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r - ||\widetilde{\mathbf{x}}_{1ij}||_2\Delta\sqrt{q/n^5}\big) - I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r\big) \\
& -P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r + ||\widetilde{\mathbf{x}}_{1ij}||_2\Delta\sqrt{q/n^5}\big) + P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r\big)\Big]
\end{aligned}
$$

where b_1 is the constant in condition (C1); and

$$
\begin{aligned}
& \widetilde{x}_{jik}\Big[I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma}\big) - I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r\big) - P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\boldsymbol{\gamma}\big) + P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}'\mathbf{t}_r\big)\Big] \\
\geq\ & 2b_1\Big[I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r + ||\widetilde{\mathbf{x}}_{1ij}||_2\Delta\sqrt{q/n^5}\big) - I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r\big) \\
& -P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r - ||\widetilde{\mathbf{x}}_{1ij}||_2\Delta\sqrt{q/n^5}\big) + P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r\big)\Big].
\end{aligned}
$$

The above positive upper bound and negative lower bound imply that

$$
\begin{aligned}
J_2 \ \leq\ & \sum_{r=1}^{N} P\Big(\max_{q+1\leq k\leq p}\sum\sum_{i\neq j} 2b_1\big[I(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r - ||\widetilde{\mathbf{x}}_{1ij}||_2\Delta\sqrt{q/n^5}) - I(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r) \\
& -P(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r + ||\widetilde{\mathbf{x}}_{1ij}||_2\Delta\sqrt{q/n^5}) + P(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r)\big] > a_1 n(n-1)\eta/8\Big) \\
+\ & \sum_{r=1}^{N} P\Big(\max_{q+1\leq k\leq p}\sum\sum_{i\neq j} 2b_1\big[-I(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r + ||\widetilde{\mathbf{x}}_{1ij}||_2\Delta\sqrt{q/n^5}) + I(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r) \\
& +P(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r - ||\widetilde{\mathbf{x}}_{1ij}||_2\Delta\sqrt{q/n^5}) - P(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r)\big] > a_1 n(n-1)\eta/8\Big) \\
=\ & J_{21} + J_{22},
\end{aligned}
$$

where the definition of J_{2i}, $i = 1, 2$, is clear from the context. To evaluate J_{21}, we have

$$
\begin{aligned}
& J_{21} \\
\leq\ & \sum_{r=1}^{N}\sum_{k=q+1}^{p} P\Big(\sum\sum_{i\neq j} 2b_1\Big[I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r - 2b_1\Delta q_n n^{-5/2}\big) - I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r\big) \\
& \quad -P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r - 2b_1\Delta q_n n^{-5/2}\big) + P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r\big)\Big] \\
& + \sum\sum_{i\neq j} 2b_1\Big[P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r - 2b_1\Delta q_n n^{-5/2}\big) - P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r + 2b_1\Delta q_n n^{-5/2}\big)\Big] \\
& \quad > a_1 n(n-1)\eta/8\Big).
\end{aligned}
$$

Note that

$$
\begin{aligned}
& \sum\sum_{i\neq j} 2b_1\Big[P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r - 2b_1\Delta q_n n^{-5/2}\big) - P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r + 2b_1\Delta q_n n^{-5/2}\big)\Big] \\
= \;& O(q_n n^{-1/2}) = o(n^2\eta).
\end{aligned}
$$

Hence for all n sufficiently large,

$$
\begin{aligned}
& J_{21} \\
\leq \;& \sum_{r=1}^{N}\sum_{k=q+1}^{p} P\Big(\sum\sum_{i\neq j} 2b_1\big[I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r - 2b_1\Delta q_n n^{-5/2}\big) - I\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r\big) \\
& \quad -P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r - 2b_1\Delta q_n n^{-5/2}\big) + P\big(\zeta_{ij} > \widetilde{\mathbf{x}}_{1ij}^T\mathbf{t}_r\big)\big] > a_1 n(n-1)\eta/16\Big) \\
= \;& o(1),
\end{aligned}
$$

following the same argument as for J_1 by applying Hoeffding's inequality. Similarly, $J_{22} \to 0$ as $n \to \infty$. This shows $D_1 \to 0$, as $n \to \infty$. The proof of the lemma is finished by noting that $||\widehat{\boldsymbol{\beta}}^{(o)} - \boldsymbol{\beta}_0||_2 = O_p(\sqrt{q/n})$. $\square$

2 Proof of Theorems 3 on the consistency of HBIC

Our proof extends the approach in Kim et al. (2016). Define

$$
\widetilde{Q}_n(\boldsymbol{\beta}) = \big[n(n-1)\big]^{-1}\sum\sum_{i\neq j}\big|(\epsilon_i - \epsilon_j) - (\mathbf{x}_i - \mathbf{x}_j)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\big|,
$$

and $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : ||\boldsymbol{\beta}||_0 \leq k_n, ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 \leq n^{-1/4}\}$.

Lemma D. *Assume the conditions of Theorem 2 are satisfied, and $k_n \log(p \vee n) = o(\sqrt{n})$. Then there exists a positive constant c such that*

$$
\begin{aligned}
P\Big(\inf_{\boldsymbol{\beta}\in\mathbb{B}} \big[\widetilde{Q}_n(\boldsymbol{\beta}) - \widetilde{Q}_n(\boldsymbol{\beta}_0) - [n(n-1)]^{-1}\sum\sum_{i\neq j}\big[I(\zeta_{ij} < 0) - \frac{1}{2}\big](\mathbf{x}_i - \mathbf{x}_j)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
-b_0||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2^2 + n^{-1/2}||\boldsymbol{\beta}||_0\big] > 0\Big) \geq 1 - 2\exp(-cn^{1/2}),
\end{aligned}
$$

for all n sufficiently large, where $b_0 = b_2 b_3$, and b_2 and b_3 are constants in conditions (C2) and (C3), respectively.

Proof. By Knight's identity,

$$\widetilde{Q}_n(\boldsymbol{\beta})-\widetilde{Q}_n(\boldsymbol{\beta}_0)-[n(n-1)]^{-1}\sum\sum_{i\neq j}\big[I(\zeta_{ij}<0)-\frac{1}{2}\big](\mathbf{x}_i-\mathbf{x}_j)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)=V_n(\boldsymbol{\beta}),$$

where $V_n(\boldsymbol{\beta})=[n(n-1)]^{-1}\sum\sum_{i\neq j}\int_0^{(\mathbf{x}_i-\mathbf{x}_j)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}[I(\zeta_{ij}<s)-I(\zeta_{ij}<0)]ds$. It is sufficient to show

$$P\Big(\inf_{\boldsymbol{\beta}\in\mathbb{B}}\big[V_n(\boldsymbol{\beta})-b_0||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2+n^{-1/2}||\boldsymbol{\beta}||_0\big]>0\Big)\geq 1-2\exp(-cn^{1/2}).\quad (8)$$

We next derive an exponential probability bound for

$$P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}}\frac{\big|V_n(\boldsymbol{\beta})-\mathrm{E}[V_n(\boldsymbol{\beta})]\big|}{||\boldsymbol{\beta}||_0}>n^{-1/2}\Big).$$

Let $M_1,\cdots,M_{m(k_n)}$ denote different subsets of $\{1,\cdots,p\}$, corresponding to different submodels with sizes at most k_n. Note that $m(k_n)\leq\binom{p}{k_n}\leq p^{k_n}$. For $l=1,\cdots,m(k_n)$, let

$$A_{M_l}=\big\{\boldsymbol{\beta}\in\mathbb{R}^p:||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2\leq n^{-1/4},\mathrm{supp}(\boldsymbol{\beta})=M_l\big\}.$$

Then $\mathbb{B}=\bigcup_{l=1}^{m(k_n)}A_{M_l}$. Let $|A_{M_l}|$ be the cardinality of A_{M_l}. For each A_{M_l}, we can cover it with l_2-balls of radius $\frac{\sqrt{|A_{M_l}|}}{8b_1}n^{-1/2}$, with centers $\boldsymbol{\beta}^*_{l_0},\cdots,\boldsymbol{\beta}^*_{l_{N_l}}$. Note that this cover can be constructed such that

$$N_l\leq\Big(\frac{2n^{-1/4}+\frac{\sqrt{|A_{M_l}|}}{8b_1}n^{-1/2}}{\frac{\sqrt{|A_{M_l}|}}{8b_1}n^{-1/2}}\Big)^{|A_{M_l}|}\leq(3n^{1/4})^{|A_{M_l}|},$$

for all n sufficiently large, assuming $k_n\ll\sqrt{n}$. For any $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2$ in the same small l_2-ball in the cover of A_{M_l},

$$\begin{aligned}
&|V_n(\boldsymbol{\beta}_1)-V_n(\boldsymbol{\beta}_2)|\\
\leq\;&[n(n-1)]^{-1}\sum\sum_{i\neq j}|(\mathbf{x}_i-\mathbf{x}_j)^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)|\\
\leq\;&2b_1||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2||_1\leq 2b_1\sqrt{|A_{M_l}|}*\frac{\sqrt{|A_{M_l}|}}{8b_1}n^{-1/2}\\
\leq\;&\frac{1}{4}|A_{M_l}|n^{-1/2}.
\end{aligned}$$

Furthermore, we note that by Lemma A in the supplementary file, $\forall\ \boldsymbol{\beta}^*_{l_j},\ j = 1, \cdots, N_l$,

$$
\begin{aligned}
& P\Big(\big|V_n(\boldsymbol{\beta}^*_{l_j}) - \mathrm{E}[V_n(\boldsymbol{\beta}^*_{l_j})]\big| > \frac{|A_{M_l}|}{2\sqrt{n}}\Big) \le 2\exp\Big[-\frac{n(n^{-1/2}|A_{M_l}|/2)^2}{8(2b_1 n^{-1/4}\sqrt{A_{M_l}})^2}\Big] \\
\le\ & 2\exp\Big(-\frac{\sqrt{n}|A_{M_l}|}{128b_1^2}\Big) \le 2\exp\Big(-\frac{\sqrt{n}}{128b_1^2}\Big).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}} \frac{\big|V_n(\boldsymbol{\beta}) - \mathrm{E}[V_n(\boldsymbol{\beta})]\big|}{||\boldsymbol{\beta}||_0} > n^{-1/2}\Big) \\
\le\ & \sum_{l=1}^{m(k_n)} P\Big(\sup_{\boldsymbol{\beta}\in A_{M_l}} \big|V_n(\boldsymbol{\beta}) - \mathrm{E}[V_n(\boldsymbol{\beta})]\big| > n^{-1/2}|A_{M_l}|\Big) \\
\le\ & 2p^{k_n}(3n^{1/4})^{k_n}\exp\Big(-\frac{\sqrt{n}}{128b_1}\Big) \\
=\ & 2\exp\Big[-\frac{\sqrt{n}}{128b_1^2} + k_n\log p + k_n\log(3n^{1/4})\Big] \\
\le\ & 2\exp(-cn^{1/2}),
\end{aligned}
$$

for all n sufficiently large, since $k_n \log p \ll \sqrt{n}$, $k_n\log(3n^{1/4}) \ll \sqrt{n}$. As a result, uniformly for $\boldsymbol{\beta}\in\mathbb{B}$,

$$V_n(\boldsymbol{\beta}) \ge \mathrm{E}[V_n(\boldsymbol{\beta})] - n^{-1/2}||\boldsymbol{\beta}||_0,$$

with probability at least $1 - 2\exp(-cn^{-1/2})$.

The proof of Theorem 1 also implies $\inf_{\boldsymbol{\beta}\in\mathbb{B}} \mathrm{E}[V_n(\boldsymbol{\beta})] \ge b_0||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2^2$. Hence, (8) is proved. $\square$

Lemma E. *Assume the conditions of Theorem 2 are satisfied, then*

$$
\begin{aligned}
& P\Big([n(n-1)]^{-1}\Big|\Big|\sum\sum_{i\neq j}\big[I(\zeta_{ij}<0) - \frac{1}{2}\big](\mathbf{x}_i - \mathbf{x}_j)\Big|\Big|_\infty > 4b_1\sqrt{\frac{\log p}{n}}\Big) \\
\le\ & 2\exp(-\log p).
\end{aligned}
$$

Proof. $\forall 1 \le k \le p,\ \forall t > 0$,

$$P\Big([n(n-1)]^{-1}\Big|\sum\sum_{i\neq j}\big[I(\zeta_{ij}<0) - \frac{1}{2}\big](x_{ik} - x_{jk})\Big| > t\Big) \le 2\exp\Big(-\frac{nt^2}{8b_1^2}\Big).$$

Hence, let $t = 4b_1\sqrt{\frac{\log p}{n}}$,

$$P\Big([n(n-1)]^{-1}\Big|\Big|\sum\sum_{i\neq j}\big[I(\zeta_{ij}<0)-\frac{1}{2}\big](\mathbf{x}_i-\mathbf{x}_j)\Big|\Big|_\infty > 4b_1\sqrt{\frac{\log p}{n}}\Big)$$
$$\leq 2p\exp\Big[-\frac{n(4b_1\sqrt{(\log p)/n})^2}{8b_1^2}\Big] = 2\exp(-\log p).$$

□

Define the following index sets:

$$\Lambda_{n-} = \{\eta > 0 : \eta \in \Lambda_n, A \not\subset A_\eta\}, \qquad \Lambda_{n+} = \{\eta > 0 : \eta \in \Lambda_n, A \subset A_\eta, A_\eta \neq A\}.$$

Given an index set S, define

$$\widehat{\boldsymbol{\beta}}_S = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p,\ \mathrm{supp}(\boldsymbol{\beta})=S} \widetilde{Q}_n(\boldsymbol{\beta}).$$

Lemma F (Uniform error bound for over-fitted values)**.** *Assume the conditions of Theorem 2 are satisfied, and that* $k_n\log(p\vee n) = o(\sqrt{n})$*. Then there exists some* $\Delta > 0$ *such that*

$$P\Big\{\sup_{S:|S|\leq k_n, S\supset A}\big(||\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_0||_2 - \Delta\sqrt{|S|(\log p)/n}\big) < 0\Big\} \to 1.$$

Proof. By the convexity of $\widetilde{Q}_n(\boldsymbol{\beta})$, it is sufficient to show

$$P\Big\{\inf_{S:|S|\leq k_n, S\supset A}\ \inf_{||\boldsymbol{\beta}_S-\boldsymbol{\beta}_0||_2=\Delta\sqrt{|S|(\log p)/n}}\big[\widetilde{Q}_n(\boldsymbol{\beta}_S)-\widetilde{Q}_n(\boldsymbol{\beta}_0)\big] > 0\Big\} \to 1. \qquad (9)$$

Since $\sqrt{k_n(\log p)/n} = o(n^{-1/4})$, then Lemma D implies that

$$\begin{aligned}
& \widetilde{Q}_n(\boldsymbol{\beta}_S)-\widetilde{Q}_n(\boldsymbol{\beta}_0)\\
\geq\ & [n(n-1)]^{-1}\sum\sum_{i\neq j}\big[I(\zeta_{ij}<0)-\frac{1}{2}\big](\mathbf{x}_i-\mathbf{x}_j)^T(\boldsymbol{\beta}_S-\boldsymbol{\beta}_0)\\
& +b_0||\boldsymbol{\beta}_S-\boldsymbol{\beta}_0||_2^2 - n^{-1/2}|S|\\
\geq\ & -[n(n-1)]^{-1}\Big|\Big|\sum\sum_{i\neq j}\big[I(\zeta_{ij}<0)-\frac{1}{2}\big](\mathbf{x}_i-\mathbf{x}_j)\Big|\Big|_\infty * ||\boldsymbol{\beta}_S-\boldsymbol{\beta}_0||_1\\
& +b_0||\boldsymbol{\beta}_S-\boldsymbol{\beta}_0||_2^2 - n^{-1/2}|S|\\
\geq\ & -4b_1\sqrt{\frac{|S|\log p}{n}} * \Delta\sqrt{\frac{|S|\log p}{n}} + b_0\Delta^2\frac{|S|\log p}{n} - n^{-1/2}|S|\\
=\ & \Delta\frac{|S|\log p}{n}(-4b_1+b_0\Delta) - n^{-1/2}|S|.
\end{aligned}$$

Take $\Delta = 3b_1b_0^{-1}$. Then (9) is proved. $\square$

Lemma G. *Assume the conditions of Theorem 2 are satisfied, and that $k_n \log(p \vee n) = o(\sqrt{n})$. Consider η_n satisfying the conditions of Theorem 2. Then*

$$P\Big\{\inf_{\eta\in\Lambda_{n+}}[HBIC(\eta) - HBIC(\eta_n)] > 0\Big\} \to 1.$$

Proof.

$$\begin{aligned}
&P\Big\{\inf_{\eta\in\Lambda_{n+}}[\mathrm{HBIC}(\eta) - \mathrm{HBIC}(\eta_n)] > 0\Big\}\\
=&P\Big\{\inf_{\eta\in\Lambda_{n+}}[\mathrm{HBIC}(\eta) - \mathrm{HBIC}(\eta_n)] > 0,\ A_{\eta_n} = A\Big\} + o(1)\\
\geq&P\Big\{\inf_{S:|S|\leq k_n, S\supset A}\Big[\log\Big(\frac{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)}{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)}\Big) + (|S|-q)\frac{\log(\log n)}{n}\log p\Big] > 0\Big\} + o(1).
\end{aligned}$$

Note that $\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S) \leq \widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)$ since S corresponds to an over-fitted model. Note that $\log(1+x) \leq x$, $\forall x > 0$, then

$$\begin{aligned}
\log\Big(\frac{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)}{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)}\Big) &= -\log\Big(\frac{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)}{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)}\Big) = -\log\Big(1 + \frac{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A) - \widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)}{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)}\Big)\\
&\geq -\frac{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A) - \widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)}{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)} \geq \frac{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S) - \widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)}{\inf_{S:|S|\leq k_n, S\supset A}\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)},
\end{aligned}$$

since $\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S) - \widetilde{Q}_n(\widehat{\boldsymbol\beta}_A) \leq 0$. By Lemma D and Lemma E, with $k_n \log p = o(\sqrt{n})$, uniformly over $\{S : |S| \leq k_n, S \supset A\}$, with probability approaching one,

$$\begin{aligned}
\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S) \geq&\widetilde{Q}_n(\boldsymbol\beta_0) + [n(n-1)]^{-1}\sum\sum_{i\neq j}\big[I(\zeta_{ij} < 0) - \frac{1}{2}\big](\mathbf{x}_i - \mathbf{x}_j)^T(\widehat{\boldsymbol\beta}_S - \boldsymbol\beta_0)\\
&+ b_0||\widehat{\boldsymbol\beta}_S - \boldsymbol\beta_0||_2^2 - n^{-1/2}||\widehat{\boldsymbol\beta}_S||_0\\
\geq&\widetilde{Q}_n(\boldsymbol\beta_0) - 4b_1\sqrt{\frac{k_n\log p}{n}} * \Delta\sqrt{\frac{k_n\log p}{n}} + b_0\Delta^2\frac{k_n\log p}{n} - n^{-1/2}k_n\\
\to&\mathrm{E}[|\epsilon_i - \epsilon_j|] \triangleq b_4.
\end{aligned}$$

Hence, with probability approaching one, $\inf_{S:|S|\le k_n, S\supset A} \widetilde{Q}_n(\widehat{\boldsymbol{\beta}}_S) \ge b_4/2$. As a result,

$$
\begin{aligned}
&P\Big\{ \inf_{\eta\in\Lambda_{n+}} [\mathrm{HBIC}(\eta) - \mathrm{HBIC}(\eta_n)] > 0 \Big\} \\
\ge & P\Big\{ \inf_{S:|S|\le k_n, S\supset A} \big[\frac{2[\widetilde{Q}_n(\widehat{\boldsymbol{\beta}}_S) - \widetilde{Q}_n(\widehat{\boldsymbol{\beta}}_A)]}{b_4} + (|S|-q)\frac{\log(\log n)}{n}\log p \big] > 0 \Big\} + o(1) \\
\ge & P\Big\{ \inf_{S:|S|\le k_n, S\supset A} \big[\frac{2[\widetilde{Q}_n(\widehat{\boldsymbol{\beta}}_S) - \widetilde{Q}_n(\boldsymbol{\beta}_0)]}{b_4} + (|S|-q)\frac{\log(\log n)}{n}\log p \big] > 0 \Big\} + o(1),
\end{aligned}
$$

since $\widetilde{Q}_n(\boldsymbol{\beta}_0) \ge \widetilde{Q}_n(\widehat{\boldsymbol{\beta}}_A)$, by the definition of $\widehat{\boldsymbol{\beta}}_A$.

Applying Lemmas D and E, we have uniformly over $\{S : |S| \le k_n, S \supset A\}$,

$$
\begin{aligned}
\widetilde{Q}_n(\widehat{\boldsymbol{\beta}}_S) - \widetilde{Q}_n(\boldsymbol{\beta}_0) \ge & [n(n-1)]^{-1} \sum\sum_{i\ne j} \big[I(\zeta_{ij} < 0) - \frac{1}{2} \big] (\mathbf{x}_i - \mathbf{x}_j)^T (\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_0) \\
& + b_0 ||\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_0||_2^2 - n^{-1/2} ||\widehat{\boldsymbol{\beta}}_S||_0 \\
\triangleq & B_n^1(S) + B_n^2(S),
\end{aligned}
$$

where

$$
\begin{aligned}
B_n^1(S) &= \mathbf{U}_{nA}^T (\widehat{\boldsymbol{\beta}}_{SA} - \boldsymbol{\beta}_{0A}) + b_0 ||\widehat{\boldsymbol{\beta}}_{SA} - \boldsymbol{\beta}_{0A}||_2^2, \\
B_n^2(S) &= \mathbf{U}_{nA^C}^T (\widehat{\boldsymbol{\beta}}_{SA^C} - \boldsymbol{\beta}_{0A^C}) + b_0 ||\widehat{\boldsymbol{\beta}}_{SA^C} - \boldsymbol{\beta}_{0A^C}||_2^2 - n^{-1/2}|S|,
\end{aligned}
$$

with $\mathbf{U}_n = [n(n-1)]^{-1} \sum\sum_{i\ne j} \big[I(\zeta_{ij} < 0) - \frac{1}{2} \big] (\mathbf{x}_i - \mathbf{x}_j)$. $\mathbf{U}_{nA}$, $\mathbf{U}_{nA^C}$, $\widehat{\boldsymbol{\beta}}_{SA}$, $\widehat{\boldsymbol{\beta}}_{SA^C}$, $\boldsymbol{\beta}_{0A}$ and $\boldsymbol{\beta}_{0A^C}$ denote the corresponding subvectors of $\mathbf{U}_n$, $\widehat{\boldsymbol{\beta}}_S$ and $\boldsymbol{\beta}_0$, according to the index set A and A^C, respectively.

Note that with probability approaching 1,

$$
\begin{aligned}
B_n^1(S) &\ge -\frac{||\mathbf{U}_{nA}||_2^2}{2b_0} \ge -\frac{q||\mathbf{U}_{nA}||_\infty^2}{2b_0} \ge -\frac{8b_1^2}{b_0} q \frac{\log q}{n}, \\
B_n^2(S) &\ge -\frac{||\mathbf{U}_{nA^C}||_2^2}{2b_0} - n^{-1/2}|S| \ge -\frac{8b_1^2}{b_0}(|S|-q)\frac{\log p}{n} - n^{-1/2}|S|.
\end{aligned}
$$

Summarizing the results above, we have

$$
\begin{aligned}
&P\Big\{\inf_{\eta\in\Lambda_{n+}}[\mathrm{HBIC}(\eta)-\mathrm{HBIC}(\eta_n)]>0\Big\}\\
\geq &P\Big\{\inf_{S:|S|\leq k_n, S\supset A}\Big[\frac{2}{b_4}\Big(-\frac{8b_1^2}{b_0}q\frac{\log q}{n}-\frac{8b_1^2}{b_0}(|S|-q)\frac{\log p}{n}-n^{-1/2}|S|\Big)\\
&\qquad\qquad +(|S|-q)\frac{\log(\log n)}{n}\log p\Big]>0\Big\}+o(1)\\
= &o(1),
\end{aligned}
$$

since $n^{-1/2}k_n=o\Big(\frac{\log(\log n)}{n}\log p\Big)$. Hence, the lemma is proved. $\square$

Lemma H. *Assume the conditions of Theorem 2 are satisfied, and that* $k_n\log(p\vee n)=o(\sqrt{n})$*. Let* $\beta^*_{\min}=\min\{|\beta_{0j}|:j\in A\}$*, and assume* $\beta^*_{\min}\gg\max\Big\{\sqrt{\frac{\log(\log n)}{n}\log p},\sqrt{\frac{q\log q}{n}}\Big\}$*. Consider* η_n *satisfying the conditions of Theorem 2. Then*

$$
P\Big\{\inf_{\eta\in\Lambda_{n-}}[HBIC(\eta)-HBIC(\eta_n)]>0\Big\}\to 1.
$$

Proof. Following the proof of Lemma G, we have

$$
\begin{aligned}
&P\Big\{\inf_{\eta\in\Lambda_{n-}}[\mathrm{HBIC}(\eta)-\mathrm{HBIC}(\eta_n)]>0\Big\}\\
=&P\Big\{\inf_{S:|S|\leq k_n, S\not\supset A}\Big[\log\Big(\frac{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)}{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)}\Big)+(|S|-q)\frac{\log(\log n)}{n}\log p\Big]>0\Big\}+o(1).
\end{aligned}
$$

Note that

$$
\begin{aligned}
\log\Big(\frac{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)}{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)}\Big)&=\log\Big(1+\frac{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)-\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)}{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)}\Big)\\
&\geq\min\Big\{\frac{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)-\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)}{2\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)},\log(2)\Big\}\\
&\geq\min\Big\{\frac{\widetilde{Q}_n(\widehat{\boldsymbol\beta}_S)-\widetilde{Q}_n(\boldsymbol\beta_0)}{2\widetilde{Q}_n(\boldsymbol\beta_0)},\log(2)\Big\},
\end{aligned}
$$

where the first inequality follows because $\log(1+x)\geq\min\{x/2,\log(2)\}$, $\forall x>0$; and the second inequality follows because $\widetilde{Q}_n(\widehat{\boldsymbol\beta}_A)\leq\widetilde{Q}_n(\boldsymbol\beta_0)$ as $\widehat{\boldsymbol\beta}_A$ is the oracle estimator.

Note that $\widetilde{Q}_n(\boldsymbol{\beta}_0) \to \mathrm{E}|\epsilon_i - \epsilon_j| \triangleq b_4$ in probability. Hence, $P\big(\widetilde{Q}_n(\boldsymbol{\beta}_0) \leq 2b_4\big) \to 1$. To prove the lemma, it is sufficient to show

$$P\Big\{ \inf_{S:|S|\leq k_n, S\not\supset A} \inf_{\boldsymbol{\beta}\in\mathbb{R}^p,\ \mathrm{supp}(\boldsymbol{\beta})=S} \Big[\frac{\widetilde{Q}_n(\boldsymbol{\beta}) - \widetilde{Q}_n(\boldsymbol{\beta}_0)}{2b_4} + (|S|-q)\frac{\log(\log n)}{n}\log p\Big] > 0\Big\} \to 1. \tag{10}$$

Consider the set

$$\widetilde{\mathbb{B}} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \mathrm{supp}(\boldsymbol{\beta}) = S,\ |S| \leq k_n, S \not\supset A,\ ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 > \Delta\sqrt{|S|(\log p)/n}\},$$

where $\Delta > 0$ is the constant in Lemma F. $\forall\ \boldsymbol{\beta} \in \widetilde{\mathbb{B}}$, write $\boldsymbol{\beta}_h = \boldsymbol{\beta}_0 + h(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, $0 < h < 1$. By the convexity of $\widetilde{Q}_n(\cdot)$, we have

$$\widetilde{Q}_n(\boldsymbol{\beta}_h) = \widetilde{Q}_n\big((1-h)\boldsymbol{\beta}_0 + h\boldsymbol{\beta}\big) \leq (1-h)\widetilde{Q}_n(\boldsymbol{\beta}_0) + h\widetilde{Q}_n(\boldsymbol{\beta}).$$

Hence, $\widetilde{Q}_n(\boldsymbol{\beta}) - \widetilde{Q}_n(\boldsymbol{\beta}_0) \geq h^{-1}\big[\widetilde{Q}_n(\boldsymbol{\beta}_h) - \widetilde{Q}_n(\boldsymbol{\beta}_0)\big]$. By definition, $||\boldsymbol{\beta}_h - \boldsymbol{\beta}_0||_2 = h||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2$. Take $h = \Delta\sqrt{|S|(\log p)/n}||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2^{-1}$, then $||\boldsymbol{\beta}_h - \boldsymbol{\beta}_0||_2 = h||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 = \Delta\sqrt{|S|(\log p)/n}$. For this choice of h, $\widetilde{Q}_n(\boldsymbol{\beta}_h) > \widetilde{Q}_n(\boldsymbol{\beta}_0)$ with probability approaching one uniformly on $\widetilde{\mathbb{B}}$. Hence,

$$P\Big(\widetilde{Q}_n(\boldsymbol{\beta}) - \widetilde{Q}_n(\boldsymbol{\beta}_0) \geq \widetilde{Q}_n(\boldsymbol{\beta}_h) - \widetilde{Q}_n(\boldsymbol{\beta}_0), \forall\ \boldsymbol{\beta} \in \widetilde{\mathbb{B}}\Big) \to 1.$$

Let

$$\mathbb{B}^* = \big\{\boldsymbol{\beta} \in \mathbb{R}^p : \mathrm{supp}(\boldsymbol{\beta}) = S,\ |S| \leq k_n, S \not\supset A,\ ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 \leq \Delta\sqrt{|S|(\log p)/n}\big\}.$$

To prove (10), it suffices to show

$$P\Big\{\inf_{\boldsymbol{\beta}\in\mathbb{B}^*} \Big[\frac{\widetilde{Q}_n(\widehat{\boldsymbol{\beta}}_S) - \widetilde{Q}_n(\boldsymbol{\beta}_0)}{2b_4} + (|S|-q)\frac{\log(\log n)}{n}\log p\Big] > 0\Big\} \to 1.$$

$\forall\ \boldsymbol{\beta} \in \mathbb{B}^*$, define $B_1 = \{j : \beta_j \neq 0, j \in A\}$, $B_2 = \{j : \beta_j = 0, j \in A\}$, $B_3 = \{j : \beta_j \neq 0, j \in A^C\}$. By Lemma D, with probability approaching one, uniformly for $\forall\ \boldsymbol{\beta} \in \mathbb{B}^*$, we have

$$\begin{aligned}
\widetilde{Q}_n(\boldsymbol{\beta}) - \widetilde{Q}_n(\boldsymbol{\beta}_0) \geq& [n(n-1)]^{-1}\sum\sum_{i\neq j}\big[I(\zeta_{ij} < 0) - \frac{1}{2}\big](\mathbf{x}_i - \mathbf{x}_j)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
&+ b_0||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2^2 - n^{-1/2}||\boldsymbol{\beta}||_0 \\
\triangleq& V_{n1} + V_{n2} + V_{n3},
\end{aligned}$$

where

$$
\begin{aligned}
V_{n1} &= \mathbf{U}_{nB_1}^T(\boldsymbol{\beta}_{B_1} - \boldsymbol{\beta}_{0B_1}) + b_0||\boldsymbol{\beta}_{B_1} - \boldsymbol{\beta}_{0B_1}||_2^2, \\
V_{n2} &= \mathbf{U}_{nB_2}^T(\boldsymbol{\beta}_{B_2} - \boldsymbol{\beta}_{0B_2}) + b_0||\boldsymbol{\beta}_{B_2} - \boldsymbol{\beta}_{0B_2}||_2^2, \\
V_{n3} &= \mathbf{U}_{nB_3}^T(\boldsymbol{\beta}_{B_3} - \boldsymbol{\beta}_{0B_3}) + b_0||\boldsymbol{\beta}_{B_3} - \boldsymbol{\beta}_{0B_3}||_2^2 - n^{-1/2}||\boldsymbol{\beta}||_0,
\end{aligned}
$$

and the notation $\mathbf{U}_{nB_j}$, $\boldsymbol{\beta}_{B_j}$ and $\boldsymbol{\beta}_{0B_j}$, $j = 1, 2, 3$, is the same as those in the proof of Lemma G.

Following the proof of Lemma D, with probability approaching one, uniformly on $\mathbb{B}^*$, we have

$$
\begin{aligned}
V_{n1} &\geq -\frac{||\mathbf{U}_{nB_1}||_2^2}{2b_0} \geq -\frac{|B_1| * ||\mathbf{U}_{nB_1}||_\infty^2}{2b_0} \geq -\frac{8b_1^2}{b_0}\frac{|B_1|\log q}{n}, \\
V_{n3} &\geq -\frac{||\mathbf{U}_{nB_3}||_2^2}{2b_0} - n^{-1/2}|S| \geq -\frac{8b_1^2}{b_0}\frac{|B_3|\log p}{n} - n^{-1/2}|S|.
\end{aligned}
$$

As the model is under-fitted, $1 \leq |B_2| \leq q$. Since $\beta^*_{\min} \gg \sqrt{\frac{q\log q}{n}}$, then

$$
\begin{aligned}
V_{n2} \geq& - ||\mathbf{U}_{nB_2}||_\infty||\boldsymbol{\beta}_{0B_2}||_1 + b_0||\boldsymbol{\beta}_{0B_2}||_2^2 \\
\geq& \big(- ||\mathbf{U}_{nB_2}||_\infty + b_0\beta^*_{\min}\big)||\boldsymbol{\beta}_{0B_2}||_1 \\
\geq& \Big(- 4b_1\sqrt{\frac{\log q}{n}} + b_0\beta^*_{\min}\Big)||\boldsymbol{\beta}_{0B_2}||_1 \geq \frac{b_0}{2}(\beta^*_{\min})^2|B_2|,
\end{aligned}
$$

for all n sufficiently large. Observing that $|B_1|+|B_3| = |S|$, $|B_1|+|B_2| = |q|$. Hence, $|S| - q = |B_3| - |B_2|$. With probability approaching one, uniformly on $\mathbb{B}^*$,

$$
\begin{aligned}
&\frac{\widetilde{Q}_n(\boldsymbol{\beta}) - \widetilde{Q}_n(\boldsymbol{\beta}_0)}{2b_4} + (|S| - q)\frac{\log(\log n)}{n}\log p \\
\geq& \frac{1}{2b_4}\Big\{ - \frac{8b_1^2}{b_0}\frac{|B_1|\log q}{n} - \frac{8b_1^2}{b_0}\frac{|B_3|\log p}{n} - n^{-1/2}(|B_1| + |B_3|) + \frac{b_0}{2}(\beta^*_{\min})^2|B_2| \\
&+ (|B_3| - |B_2|)\frac{\log(\log n)}{n}\log p\Big\} \\
\geq& \frac{1}{2b_4}\Big\{ - q\Big(\frac{8b_1^2\log q}{b_0 n} + n^{-1/2}\Big) + |B_2|\Big(\frac{b_0}{2}(\beta^*_{\min})^2 - \frac{\log(\log n)}{n}\log p\Big) \\
&+ |B_3|\Big(- \frac{8b_1^2\log p}{b_0 n} - n^{-1/2} + \frac{\log(\log n)}{n}\log p\Big)\Big\} \\
>& 0,
\end{aligned}
$$

for all n sufficiently large, since $\beta^*_{\min} \gg \max\left\{\sqrt{\frac{\log(\log n)}{n}}\log p, \sqrt{\frac{q\log q}{n}}, \sqrt{k_n n^{-1/2}}\right\}$. Note that $q \le k_n$, $|B_3| \le k_n$ and $k_n n^{-1/2} = o(1)$. The lemma is proved. $\square$

Proof of Theorem 3. It follows immediately by combining the results of Lemma G and Lemma H. $\square$

3 Additional numerical results

This section reports the results from additional simulation studies.

Example S1. We consider the same data generative model as in Example 1 in the main paper except that $\boldsymbol{\beta}_0 = (2, 1.5, 1.25, 1, 0.75, 0.5, 0.25, \mathbf{0}_{p-7})^T$, where $\mathbf{0}_{p-7}$ is a $(p-7)$-dimensional vector of zeros. Comparing with Example 1, this is a more challenging scenario with 7 active variables and some weaker signals. Table S1 summarizes the simulations results, which showed similar performance as in Example 1.

Example S2. We consider the same data generative model as in Example 1 in the main paper except that $\boldsymbol{\beta}_0 = (2, 2, 2, 1.5, 1.5, 1.25, 1.25, 1, 1, 0.75, 0.75, 0.5, 0.5, 0.25, 0.25, \mathbf{0}_{p-15})^T$, where $\mathbf{0}_{p-15}$ is a $(p-15)$-dimensional vector of zeros. Comparing with Example 1, this is a considerably more challenging scenario with 15 active variables and more weaker signals. Table S2 summarizes the simulations results, which demonstrate similar performance as in Example 1.

Example S3. We consider the same data generative model with $N(0, 1)$ error as in Example 1 and investigate the effect of different choices of n and p. Table S3 summarizes the simulations results. We observe similar performance as in Example 1.

Example S4. This example provides more information on the simulated tuning parameter. Figure S1 depicts the histogram of $c||\mathbf{S}_n||_\infty$ for the simulation setup in Example 1 with $N(0, 1)$ error and $c = 1.01$. For a given small $\alpha_0 > 0$, the $(1-\alpha)$-quantile of $c||\mathbf{S}_n||_\infty$ falls below 0.4. The theoretical upper bound of $c||\mathbf{S}_n||_\infty$ given in Lemma 2 of the main paper shows that the simulated tuning parameter is of order $O(\sqrt{\log p/n})$ with high probability. However, this upper bound is not expected to be sharp. In the setting of Example 1, this bound is around 2.5$\sim$3. We observe that using this bound

Table S1: Simulation results for Example S1

Error	Method	L1 error	L2 error	ME	FP	FN
$N(0, 0.25)$	Lasso	1.67 (0.02)	0.42 (0.00)	0.09 (0.00)	24.3 (0.48)	0 (0)
	$\sqrt{\text{Lasso}}$	1.48 (0.02)	0.41 (0.00)	0.10 (0.00)	18.64 (0.31)	0 (0)
	SCAD	0.58 (0.02)	0.26 (0.01)	0.04 (0.00)	0.36 (0.04)	0.28 (0.03)
	Rank Lasso	1.69 (0.02)	0.48 (0.01)	0.17 (0.00)	17.86 (0.36)	0 (0)
	Rank SCAD	1.03 (0.02)	0.39 (0.01)	0.09 (0.00)	3.71 (0.20)	0.58 (0.03)
$N(0, 1)$	Lasso	3.31 (0.06)	0.83 (0.01)	0.38 (0.01)	22.63 (0.44)	0.34 (0.04)
	$\sqrt{\text{Lasso}}$	2.94 (0.04)	0.80 (0.01)	0.37 (0.01)	18.72 (0.32)	0.38 (0.04)
	SCAD	1.51 (0.03)	0.64 (0.01)	0.23 (0.01)	0.41 (0.04)	1.38 (0.04)
	Rank Lasso	3.29 (0.05)	0.93 (0.01)	0.65 (0.01)	17.3 (0.37)	0.67 (0.04)
	Rank SCAD	1.90 (0.03)	0.70 (0.01)	0.28 (0.01)	5.04 (0.25)	1 (0)
$N(0, 2)$	Lasso	4.72 (0.08)	1.18 (0.02)	0.76 (0.02)	24.87 (0.63)	0.68 (0.04)
	$\sqrt{\text{Lasso}}$	4.20 (0.06)	1.13 (0.01)	0.76 (0.02)	18.58 (0.26)	0.68 (0.04)
	SCAD	2.24 (0.06)	0.93 (0.02)	0.47 (0.02)	0.50 (0.05)	1.79 (0.05)
	Rank Lasso	4.55 (0.07)	1.28 (0.02)	1.27 (0.03)	16.72 (0.31)	1 (0)
	Rank SCAD	2.96 (0.06)	0.99 (0.01)	0.56 (0.02)	7.75 (0.43)	1.52 (0.05)
MN	Lasso	6.35 (0.20)	1.63 (0.05)	1.64 (0.08)	20.23 (0.34)	1.44 (0.09)
	$\sqrt{\text{Lasso}}$	5.82 (0.17)	1.60 (0.05)	1.64 (0.07)	17.89 (0.29)	1.32 (0.07)
	SCAD	4.12 (0.19)	1.53 (0.06)	1.47 (0.09)	1.64 (0.12)	2.62 (0.07)
	Rank Lasso	0.59 (0.01)	0.22 (0.01)	0.05 (0.00)	18.09 (0.36)	0 (0)
	Rank SCAD	0.42 (0.01)	0.21 (0.01)	0.04 (0.00)	0.92 (0.08)	0 (0)
$\sqrt{2}t_4$	Lasso	6.17 (0.10)	1.57 (0.02)	1.38 (0.03)	22.29 (0.44)	1.27 (0.06)
	$\sqrt{\text{Lasso}}$	5.60 (0.08)	1.53 (0.02)	1.40 (0.03)	18.09 (0.28)	1.31 (0.06)
	SCAD	3.91 (0.09)	1.53 (0.03)	1.31 (0.05)	1.41 (0.09)	2.68 (0.05)
	Rank Lasso	5.40 (0.08)	1.54 (0.02)	1.90 (0.04)	16.68 (0.35)	1.52 (0.06)
	Rank SCAD	3.73 (0.08)	1.23 (0.02)	0.87 (0.02)	7.02 (0.37)	1.76 (0.06)
Cauchy	Lasso	12.71 (0.22)	3.94 (0.08)	14.59 (0.91)	9.14 (0.55)	5.84 (0.10)
	$\sqrt{\text{Lasso}}$	11.67 (0.19)	3.75 (0.08)	15.23 (0.93)	7.20 (0.40)	5.90 (0.10)
	SCAD	8.89 (0.09)	3.48 (0.04)	34.29 (0.32)	0.00 (0.00)	7.00 (0.00)
	Rank Lasso	8.02 (0.15)	2.41 (0.05)	5.36 (0.27)	15.15 (0.30)	2.15 (0.07)
	Rank SCAD	6.89 (0.19)	2.41 (0.07)	4.24 (0.26)	5.05 (0.24)	2.58 (0.07)

directly as the tuning parameter leads to over-penalization.

Table S2: Simulation results for Example S2

Error	Method	L1 error	L2 error	ME	FP	FN
$N(0, 0.25)$	Lasso	3.71 (0.04)	0.73 (0.01)	0.31 (0.01)	34.15 (0.32)	0 (0)
	$\sqrt{\text{Lasso}}$	3.81 (0.04)	0.74 (0.01)	0.29 (0.01)	35.77 (0.35)	0 (0)
	SCAD	1.39 (0.03)	0.44 (0.01)	0.10 (0.00)	0.48 (0.05)	1.03 (0.05)
	Rank Lasso	4.34 (0.04)	0.86 (0.01)	0.47 (0.01)	36.58 (0.35)	0 (0)
	Rank SCAD	2.35 (0.03)	0.59 (0.01)	0.19 (0.00)	14.46 (0.51)	0.64 (0.05)
$N(0, 1)$	Lasso	7.31 (0.08)	1.37 (0.01)	0.96 (0.02)	38.68 (0.36)	1.31 (0.06)
	$\sqrt{\text{Lasso}}$	7.11 (0.08)	1.36 (0.01)	0.99 (0.02)	35.03 (0.33)	1.23 (0.06)
	SCAD	3.88 (0.07)	1.13 (0.02)	0.69 (0.02)	1.76 (0.12)	3.30 (0.06)
	Rank Lasso	8.07 (0.07)	1.60 (0.01)	1.67 (0.03)	35.52 (0.36)	0.60 (0.05)
	Rank SCAD	4.78 (0.07)	1.16 (0.01)	0.72 (0.02)	14.71 (0.56)	2.12 (0.08)
$N(0, 2)$	Lasso	10.12 (0.11)	1.84 (0.02)	1.76 (0.03)	40.3 (0.51)	2.34 (0.06)
	$\sqrt{\text{Lasso}}$	9.54 (0.09)	1.83 (0.01)	1.81 (0.03)	34.24 (0.34)	2.30 (0.06)
	SCAD	5.95 (0.10)	1.68 (0.02)	1.49 (0.04)	2.26 (0.13)	4.49 (0.06)
	Rank Lasso	11.06 (0.10)	2.19 (0.02)	3.09 (0.05)	35.94 (0.29)	1.93 (0.08)
	Rank SCAD	7.40 (0.12)	1.76 (0.02)	1.66 (0.04)	16.09 (0.57)	3.10 (0.09)
MN	Lasso	13.51 (0.32)	2.56 (0.06)	3.61 (0.14)	34.52 (0.44)	3.44 (0.13)
	$\sqrt{\text{Lasso}}$	12.94 (0.28)	2.53 (0.05)	3.67 (0.15)	31.89 (0.40)	3.44 (0.13)
	SCAD	8.80 (0.34)	2.34 (0.08)	3.27 (0.18)	3.36 (0.19)	5.42 (0.17)
	Rank Lasso	2.06 (0.04)	0.43 (0.01)	0.14 (0.01)	41.80 (0.44)	0 (0)
	Rank SCAD	0.74 (0.02)	0.25 (0.01)	0.05 (0.00)	3.26 (0.14)	0 (0)
$\sqrt{2}t_4$	Lasso	12.92 (0.16)	2.38 (0.02)	2.97 (0.06)	38.24 (0.60)	3.15 (0.08)
	$\sqrt{\text{Lasso}}$	12.10 (0.14)	2.33 (0.02)	2.96 (0.06)	32.61 (0.36)	3.10 (0.08)
	SCAD	8.27 (0.16)	2.29 (0.04)	2.78 (0.09)	2.83 (0.15)	5.53 (0.07)
	Rank Lasso	13.42 (0.12)	2.67 (0.02)	4.74 (0.08)	34.66 (0.37)	2.93 (0.11)
	Rank SCAD	9.85 (0.16)	2.32 (0.03)	2.96 (0.09)	14.15 (0.46)	4.21 (0.12)
Cauchy	Lasso	27.51 (0.39)	6.27 (0.14)	23.79 (1.08)	17.99 (0.82)	11.51 (0.23)
	$\sqrt{\text{Lasso}}$	25.74 (0.35)	5.90 (0.12)	24.73 (1.16)	15.13 (0.64)	11.62 (0.23)
	SCAD	19.85 (0.14)	5.17 (0.05)	108.04 (4.85)	6.34 (0.55)	13.03 (0.19)
	Rank Lasso	19.85 (0.26)	4.06 (0.06)	12.30 (0.39)	31.32 (0.32)	6.14 (0.15)
	Rank SCAD	18.37 (0.35)	4.61 (0.09)	12.76 (0.50)	9.03 (0.28)	8.52 (0.16)

Example S5. Table S4 below summarizes the simulations results for the setup in Example 1 with N(0,1) error and $c = 1.1$.

Example S6 (Huber loss). In this example, we compare the proposed methods with high-dimensional penalized regression with the Huber loss function (Huber et al., 1964). Huber's loss function played an important role in classical robust statistics. Recently, several papers have studied the theory of high-dimensional Huber's regression, see for instance Fan et al.

Table S3: Simulation results for Example S3

n	p	Method	L1 error	L2 error	ME	FP	FN
50	400	Lasso	2.76 (0.07)	0.88 (0.01)	0.44 (0.01)	16.41 (0.65)	0 (0)
		$\sqrt{\text{Lasso}}$	2.10 (0.04)	0.84 (0.01)	0.46 (0.01)	8.90 (0.22)	0 (0)
		SCAD	0.75 (0.03)	0.44 (0.02)	0.14 (0.01)	0.36 (0.04)	0 (0)
		Rank Lasso	2.17 (0.04)	1.01 (0.02)	0.96 (0.03)	5.65 (0.20)	0 (0)
		Rank SCAD	1.37 (0.06)	0.55 (0.02)	0.20 (0.01)	6.28 (0.60)	0 (0)
100	400	Lasso	1.73 (0.04)	0.60 (0.01)	0.21 (0.01)	14.18 (0.44)	0 (0)
		$\sqrt{\text{Lasso}}$	1.46 (0.03)	0.58 (0.01)	0.22 (0.01)	10.30 (0.29)	0 (0)
		SCAD	0.46 (0.02)	0.27 (0.01)	0.05 (0.00)	0 (0)	0 (0)
		Rank Lasso	1.57 (0.03)	0.67 (0.01)	0.38 (0.01)	8.56 (0.23)	0 (0)
		Rank SCAD	0.47 (0.01)	0.26 (0.01)	0.05 (0.00)	0.23 (0.04)	0 (0)
100	1000	Lasso	2.50 (0.04)	0.81 (0.01)	0.36 (0.01)	21.15 (0.48)	0 (0)
		$\sqrt{\text{Lasso}}$	2.21 (0.04)	0.78 (0.01)	0.35 (0.01)	16.67 (0.33)	0 (0)
		SCAD	0.44 (0.02)	0.27 (0.01)	0.05 (0.00)	0 (0)	0 (0)
		Rank Lasso	2.23 (0.03)	0.86 (0.01)	0.53 (0.01)	13.80 (0.32)	0 (0)
		Rank SCAD	0.44 (0.01)	0.25 (0.01)	0.05 (0.00)	0 (0)	0 (0)
100	5000	Lasso	2.87 (0.13)	0.84 (0.03)	0.39 (0.02)	27.06 (1.28)	0 (0)
		$\sqrt{\text{Lasso}}$	2.33 (0.08)	0.79 (0.02)	0.34 (0.02)	21.43 (0.84)	0 (0)
		SCAD	0.42 (0.01)	0.24 (0.01)	0.04 (0.00)	0 (0)	0 (0)
		Rank Lasso	2.04 (0.07)	0.73 (0.02)	0.32 (0.02)	18.28 (0.98)	0 (0)
		Rank SCAD	0.47 (0.02)	0.26 (0.01)	0.05 (0.00)	0 (0)	0 (0)

Table S4: Simulation results for Example 1 with N(0,1) error and $c = 1.1$

Method	L1 error	L2 error	ME	FP	FN
Lasso	1.54 (0.04)	0.57 (0.01)	0.20 (0.01)	13.08 (0.43)	0 (0)
$\sqrt{\text{Lasso}}$	1.41 (0.03)	0.54 (0.01)	0.21 (0.01)	11.46 (0.33)	0 (0)
SCAD	0.46 (0.02)	0.28 (0.01)	0.06 (0.00)	0 (0)	0 (0)
Rank Lasso	0.92 (0.01)	0.50 (0.01)	0.32 (0.01)	2.80 (0.15)	0 (0)
Rank SCAD	0.48 (0.01)	0.28 (0.01)	0.06 (0.00)	0 (0)	0 (0)

(2017), Loh (2017) and Sun et al. (2020). In this example, we adopt the same simulation setup as in Section 5 of Sun et al. (2020) where $\epsilon_i \sim t_{1.5}$ and $p = 500$. The L_2 estimation error of the proposed methods are summarized in Table S5 and displayed in Figure S2. Comparing the results with what were shown in Figure 2 of Sun et al. (2020), we observe that the proposed new methods have improved performance in estimation accuracy. For example, for $n = 400$, the L_2 estimation error for Rank Lasso is 0.58 and that for

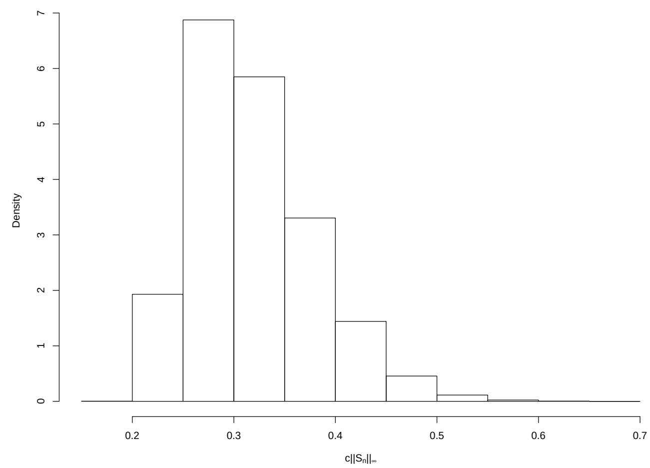

Figure S1: Histogram of $c||\mathbf{S}_n||_\infty$ in Example 1 for $N(0,1)$ error

Rank SCAD is 0.30; in contrast, the L_2 estimation error in Figure 2 of Sun et al. (2020) is close to 1. In Sun et al. (2020), the tuning parameter of Huber loss is set as $\tau = \frac{\widehat{\sigma}}{2}\sqrt{\frac{n}{\log p \log n}}$, and $\widehat{\sigma} = n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$.

Table S5: Simulation results for Example S6: L_2 estimation error

n	Rank Lasso	Rank SCAD
100	1.95 (0.01)	0.68 (0.06)
200	0.94 (0.01)	0.45 (0.03)
300	0.70 (0.02)	0.36 (0.03)
400	0.58 (0.01)	0.30 (0.02)
500	0.51 (0.02)	0.27 (0.02)

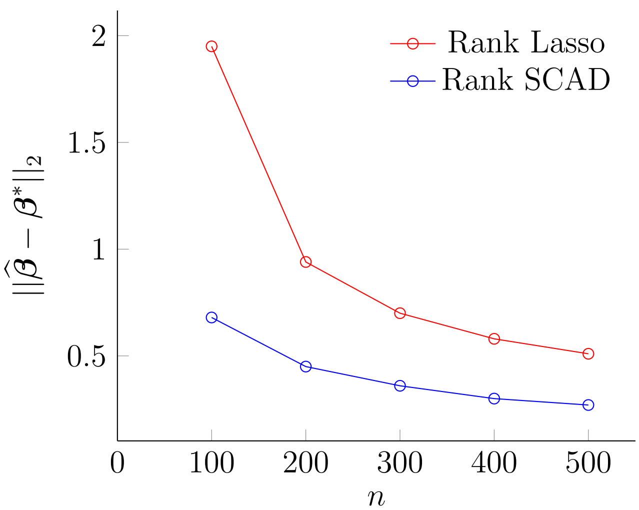

Figure S2: L_2 estimation error of the proposed methods for Example 7

Example S7. In this example, we consider the simulation setting of Example 3 of the main paper with the mixture normal random error. As in the main paper, we consider cross-validated Lasso (with tuning parameter minimizing the cross-validation error), square-root Lasso, Huber Lasso and Rank Lasso. Upon a referee's suggestion, we also include Lasso-1se and Lasso-EBIC (Lasso with tuning parameter selected by the extended BIC (EBIC) of Chen and Chen (2008)). Lasso-1se is cross-validated Lasso with tuning parameter selected by the one standard error rule (an option in "cv.glmnet" function), see for example Section 2.3 of Hastie et al. (2015). Lasso-1se is implemented in the R package glmnet, while Lasso-EBIC is implemented using the R package SIS (Saldana and Feng (2018)). The results are summarized in Table S6. The tuning parameters for Huber Lasso are selected via a two-dimensional grid search. We observe that both Huber Lasso and Rank Lasso substantially improve the performance of Lasso (with different tuning parameter selection methods) and square-root Lasso in the presence the heavy-tailed errors. For all three different design matrices, Rank Lasso yields the smallest estimation error.

Table S6: Simulation results for Example S7

Σ	Method	L1 error	L2 error	Pred error	FP	FN
Σ_1	Lasso	4.87 (0.18)	1.75 (0.06)	0.76 (0.04)	12.30 (0.25)	0 (0)
	$\sqrt{\text{Lasso}}$	4.79 (0.17)	1.72 (0.06)	0.75 (0.04)	12.23 (0.26)	0 (0)
	Lasso-1se	3.88 (0.13)	1.78 (0.05)	3.56 (0.18)	5.24 (0.21)	0.33 (0.04)
	Lasso-EBIC	3.99 (0.14)	1.64 (0.05)	1.32 (0.05)	6.54 (0.30)	0 (0)
	Huber Lasso	1.28 (0.04)	0.50 (0.01)	0.11 (0.01)	10.77 (0.23)	0 (0)
	Rank Lasso	0.27 (0.01)	0.11 (0.00)	0.00 (0.00)	12.29 (0.27)	0 (0)
Σ_2	Lasso	2.49 (0.10)	0.92 (0.03)	0.82 (0.05)	11.37 (0.35)	0 (0)
	$\sqrt{\text{Lasso}}$	2.18 (0.08)	0.93 (0.03)	0.81 (0.05)	7.65 (0.23)	0 (0)
	Lasso-1se	2.71 (0.08)	1.50 (0.04)	3.35 (0.17)	0 (0)	0 (0)
	Lasso-EBIC	1.90 (0.06)	1.01 (0.03)	1.19 (0.05)	1.40 (0.13)	0 (0)
	Huber Lasso	0.49 (0.01)	0.27 (0.01)	0.08 (0.00)	2.47 (0.16)	0 (0)
	Rank Lasso	0.11 (0.00)	0.07 (0.00)	0.01 (0.00)	1.62 (0.11)	0 (0)
Σ_3	Lasso	1.53 (0.06)	0.67 (0.02)	0.59 (0.04)	7.12 (0.33)	0 (0)
	$\sqrt{\text{Lasso}}$	1.55 (0.06)	0.68 (0.02)	0.57 (0.04)	6.32 (0.19)	0 (0)
	Lasso-1se	2.17 (0.06)	1.24 (0.03)	2.71 (0.12)	0 (0)	0 (0)
	Lasso-EBIC	1.17 (0.04)	0.67 (0.02)	0.62 (0.03)	0.84 (0.13)	0 (0)
	Huber Lasso	0.31 (0.01)	0.19 (0.01)	0.05 (0.00)	0.31 (0.04)	0 (0)
	Rank Lasso	0.08 (0.00)	0.05 (0.00)	0.00 (0.00)	0 (0)	0 (0)

References

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Fan, J., Li, Q., and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B*, 79(1):247–265.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Huber, P. J. et al. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Kim, Y., Jeon, J.-J., et al. (2016). Consistent model selection criteria for quadratically supported risks. *The Annals of Statistics*, 44(6):2467–2496.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, (7):186–199.

Saldana, D. F. and Feng, Y. (2018). SIS: An R package for sure independence screening in ultrahigh-dimensional statistical models. *Journal of Statistical Software*, 83(2):1–25.

Sun, Q., Zhou, W.-X., and Fan, J. (2020). Adaptive huber regression. *Journal of the American Statistical Association*, 115:254–265.